



Zhang, F., Katsenou, A. V., Afonso, M., Dimitrov, G., & Bull, D. R. (2020). Comparing VVC, HEVC and AV1 using Objective and Subjective Assessments. Unpublished.
<https://arxiv.org/abs/2003.10282>

Early version, also known as pre-print

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the submitted manuscript (SM). It first appeared online via arXiv at <https://arxiv.org/abs/2003.10282>. Please refer to any applicable terms of use of the author.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Comparing VVC, HEVC and AV1 using Objective and Subjective Assessments

Fan Zhang, *Member, IEEE*, Angeliki V. Katsenou, *Member, IEEE*, Mariana Afonso, *Member, IEEE*, Goce Dimitrov, and David R. Bull, *Fellow, IEEE*

Abstract—In this paper, the performance of three state-of-the-art video codecs: High Efficiency Video Coding (HEVC) Test Model (HM), AOMedia Video 1 (AV1) and Versatile Video Coding Test Model (VTM), are evaluated using both objective and subjective quality assessments. Nine source sequences were carefully selected to offer both diversity and representativeness, and different resolution versions were encoded by all three codecs at pre-defined target bitrates. The compression efficiency of the three codecs are evaluated using two commonly used objective quality metrics, PSNR and VMAF. The subjective quality of their reconstructed content is also evaluated through psychophysical experiments. Furthermore, HEVC and AV1 are compared within a dynamic optimization framework (convex hull rate-distortion optimization) across resolutions with a wider bitrate, using both objective and subjective evaluations. Finally the computational complexities of three tested codecs are compared. The subjective assessments indicate that, for the tested versions there is no significant difference between AV1 and HM, while the tested VTM version shows significant enhancements. The selected source sequences, compressed video content and associated subjective data are available online, offering a resource for compression performance evaluation and objective video quality assessment.

Index Terms—Codec comparison, HEVC, AV1, VVC, dynamic optimizer, objective quality assessment and subjective quality assessment.

I. INTRODUCTION

Video technology is ubiquitous in modern life, with wired and wireless video streaming, terrestrial and satellite TV, Blu-ray players, digital cameras, video conferencing and surveillance all underpinned by efficient signal representations. It has been predicted that, by 2022, 82% (approximately 4.0ZB) of all global internet traffic per year will be video content [1]. It is therefore a very challenging time for compression, which must efficiently code these increased quantities of video at higher spatial and temporal resolutions, dynamic resolutions and qualities.

The last three decades have witnessed significant advances in video compression technology, from the first international video coding standard H.120 [2], to the widely adopted MPEG-2/H.262 [3] and H.264/AVC (Advanced Video Coding)

[4] standards. More recently, ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) have initiated the development of a new video coding standard, Versatile Video Coding (VVC) [5], with the aim of reducing bit rates by 30%-50% compared to the current High Efficiency Video Coding (HEVC) standard [6]. In parallel, the Alliance for Open Media (AOMedia) has developed royalty-free open-source video codecs to compete with MPEG standards. The recently launched AOMedia Video 1 (AV1) codec [7] has been reported to outperform its predecessor VP9 [8].

In order to benchmark these coding algorithms, their rate quality performance can be evaluated using objective and/or subjective assessment methods. Existing work [9–14] has reported comparisons for contemporary codecs, but the results have been varied and the conclusions confusing due to the use of different coding configurations.

In this context, this paper presents a comparison between the test models for three major video coding standards (HEVC, AV1 and VVC) using their corresponding common test conditions to create a fair comparison. The results are based on eighteen representative source sequences at UHD (3840×2160) and HD (1920×1080) resolutions using traditional (constant resolution) and Dynamic Optimizer (DO)¹ (DO) [15] (for HD resolution only) approaches.

We provide a comprehensive extension of our previous work in [16], where only AV1 and HEVC comparison results were presented based on the DO approach. Comparing to existing work on codec comparison [9, 10, 12, 13], this paper is the first to present objective and subjective comparison results for the VVC Test Model and to compare codecs within an adaptive streaming framework.

In this paper, three specific research questions are addressed:

- 1) What is the overall compression efficiency of the tested video codecs in terms of subjective and objective video quality?
- 2) How does the compression efficiency vary across various bit rates and resolutions?
- 3) How does the performance of commonly used objective quality metrics correlate with collected subjective scores?
- 4) What are the computational complexity figures for the tested codecs?

The rest of this paper is organised as follows. Section II briefly reviews the history of video coding and related

Manuscript drafted at January 2020.

The authors acknowledge funding from the UK Engineering and Physical Sciences Research Council (EPSRC, project No. EP/M000885/1), the Leverhulme Early Career Fellowship, and the support from NVIDIA Corporation for the donation of GPUs.

Fan Zhang, Angeliki V. Katsenou, Mariana Afonso, Goce Dimitrov, and David R. Bull are with the Bristol Vision Institute, University of Bristol, Bristol, UK.

E-mail: {fan.zhang, angeliki.katsenou, mariana.afonso, goce.dimitrov, dave.bull}@bristol.ac.uk

¹Here only convex hull rate-distortion optimisation within each shot is employed.

work on codec comparison. Section III presents the selected source sequences and the coding configurations employed in generating various compressed content. In Section IV, the conducted subjective experiments are described in detail, while the comparison results through both objective and subjective assessment are reported and discussed in Section V. Finally, Section VI outlines the conclusion and future work.

II. BACKGROUND

This section provides a brief overview of video coding standards and summarises previous work on video codec comparisons.

A. Video coding standards

Video coding standards normally define the syntax of bitstream and the decoding process, while encoders generate standard-compliant bitstream and thus determine compression performance. Each generation of video coding standard comes with a reference test model, such as HM (HEVC Test Model) for HEVC, which can be used to provide a performance benchmark.

H.264/MPEG-4-AVC [4] was launched in 2004, and is still the most prolific video coding standard, despite the fact that the current standard, H.265/HEVC [6] finalised in 2013, provides enhanced coding performance. Since 2018, work on the next generation video coding standard, Versatile Video Coding (VVC), has targeted 30%-50% coding gain over H.265/HEVC, supporting immersive formats (360° videos) and higher spatial resolutions, up to 16K.

B. Other video coding technologies

Alongside recent MPEG standardisation, there has been increasing activity in the development of open-source royalty-free video codecs, particularly by the Alliance for Open Media (AOMedia), a consortium of video-related companies. VP9 [8] was developed by Google to compete with MPEG and provided a basis for AV1 (AOMedia Video 1) [7] which was released in 2018. AV1 is expected to be a primary competitor for the current MPEG video coding standards, especially in the context of streaming applications.

For further details on existing video coding standards and formats, the readers are referred to [17–19].

C. Codec comparison

The performance of video coding algorithms is usually assessed by comparing their rate-distortion (RD) or rate-quality (RQ) performance on various test sequences. The selection of test content is important and should provide a diverse and representative coverage of the video parameter space. Objective quality metrics or subjective opinion measurements are normally employed to assess compressed video quality, and the overall RD or RQ performance difference between codecs can be then calculated using Bjøntegaard measurements (for objective quality metrics) [20] or SCENIC [21] (for subjective assessments). Recently, in order to compare video codecs and

optimise rate quality performance, the DO approach, particularly its convex hull rate-distortion optimisation, has been developed by Netflix [15] for adaptive streaming applications. This constructs a convex hull over rate quality curves at various spatial resolutions, and provides a fairer approach for comparing different codecs across a wider bit rate range and resolutions.

Most recent work has focused on comparisons between MPEG codecs (H.264/AVC and HEVC) and royalty-free (VP9 and AV1) codecs [9–11, 22] and on their application in adaptive steaming services [12–14]. However the results presented are acknowledged to be highly inconsistent, mainly due to the different configurations employed. Moreover, as far as we are aware, there have been no subjective codec comparisons in the context of adaptive streaming or including the performance VVC.

III. TEST CONTENT AND CODEC CONFIGURATIONS

This section describes the selection of source sequences and the different codec configurations used to generate their various compressed versions.

A. Source sequence selection

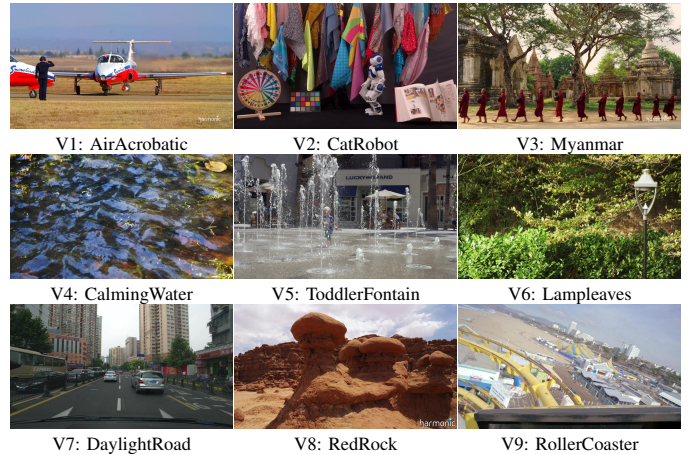


Fig. 1: Sample video frames from the selected source sequences.

Nine source sequences were selected from Harmonic [26], BVI-Texture and JVT (Joint Video Exploration Team) CTC (Common Test Conditions) datasets. Each sequence is progressively scanned, at Ultra High Definition (UHD, 3840×2160) resolution, with a frame rate of 60 frames per second (fps), and without scene-cuts. All were truncated from their original lengths to five seconds (rather than the recommended 10 seconds in ITU standard [27]). This reflects the recommendations of a recent study on optimal video duration for quality assessment [28, 29]. Sample frames from the selected nine source sequences alongside clip names and indices are shown in Fig. 1. The dataset includes three sequences with only local motion (without any camera motion, V1-V3), three sequences with dynamic textures (for definitions see [30], V4-V6), and three with complex camera movements (V7-V9). The coverage of the video parameter space is confirmed in Fig. 2, where

TABLE I: The software versions and configurations of the evaluated video codecs.

Codec	Version	Configuration parameters
HEVC HM	16.18	Random access configuration for Main10 profile [23]. IntraPeriod=64 and GOPSize=16.
AOM AV1	0.1.0-9647-ga6fa0877f	Common settings with high latency CQP configuration [24]. Other coding parameters: passes=2, cpu-used=1, kf-max-dist=64, kf-min-dist=64, arnr-maxframes=7, arnr-strength=5, lag-in-frames=16, aq-mode=0, bias-pct=100, minsection-pct=1, maxsection-pct=10000, auto-alt-ref=1, min-q=0, max-q=63, max-gf-interval=16, min-gf-interval=4 and color-primaries=bt709.
VVC VTM	4.01	Random access configuration [25]. IntraPeriod=64 and GOPSize=16.

the Spatial and Temporal Information of the dataset (SI and TI) [31] are plotted.

In order to investigate coding performance for different resolutions and within an adaptive streaming framework, three spatial resolution groups were generated from the source sequences: (A) UHD (3840×2160) only, (B) HD (1920×1080) only, and (C) HD-Dynamic Optimizer (HD-DO). For group C, coding results for three different resolutions (1920×1080 , 1280×720 , and 960×544) and with various quantisation parameters (QPs) were firstly generated. The reconstructed videos were then up-sampled to HD resolution (in order to provide a basis for comparison with the original HD sequences). Here, spatial resolution re-sampling was implemented using Lanczos-3 filters [32]. The rate points with optimal rate-quality performance (based on VMAF [33]) were selected across the three tested resolutions for each target bit rate and codec. This process is repeated to create the entire convex hull in the DO approach [15].

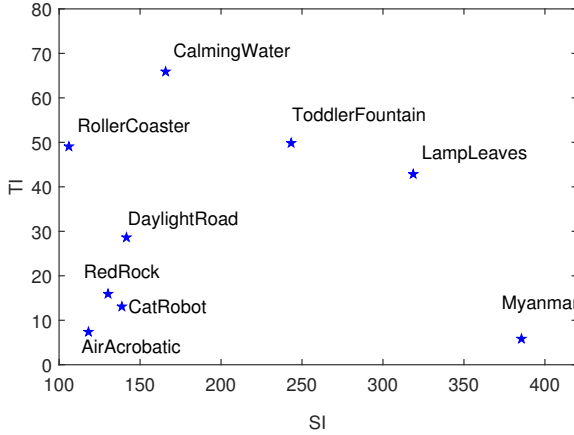


Fig. 2: Scatter plot of SI and TI for the selected source sequences.

B. Coding configurations

The reference test models of HEVC and VVC, and their major competitor, AV1 have been evaluated in this study. Each codec was configured using the coding parameters defined in their common test conditions [23–25], with fixed quantisation parameters (rate control disabled), the same structural delay (e.g. defined as GOP size in the HEVC HM software) of 16 frames and the same random access intervals (e.g. defined as IntraPeriod in the HEVC HM software) of 64 frames.

The actual codec versions and configuration parameters are provided in Table I.

Different target bit rates were pre-determined for each test sequence and for each resolution group (four points for resolution group A and B, and five for HD-DO group), and their values are shown in Table II. These were determined based on the preliminary encoding results of the test sequences for each resolution group using AV1. This decision was made because the version of AV1 employed restricted production of bitstreams at pre-defined bit rates, as only integer quantisation parameters could be used. On the other hand, for HEVC HM and VVC VTM this was easier to achieve by enabling the “QPIncrementFrame” parameter. In order to achieve these target bitrates, the quantisation parameter values were iteratively adjusted to ensure the output bit rates were sufficiently close to the targets (within a range of $\pm 3\%$).

C. Summary

In summary, a total number of 306 distorted sequences were produced: there are 108 (9 source sequences \times 4 rate points \times 3 codecs) for Resolution Group A (UHD only), 108 ($9 \times 4 \times 3$) for Resolution Group B (HD only), and 90 ($9 \times 5 \times 2$) for Resolution Group C (HD-DO)².

IV. SUBJECTIVE EXPERIMENTS

Three subjective experiment sessions were conducted separately on the test sequences in the three resolution groups. The experimental setup, procedure, test methodology and data processing approach are reported in this section.

A. Environmental Setup

All three experiment sessions were conducted in a darkened, living room-style environment. The background luminance level was set to 15% of the peak luminance of the monitor used (62.5 lux) [27]. All test sequences were shown at their native spatial resolution and frame rates, on a consumer display, a SONY KD65Z9D LCD TV, which measures 1429×804 mm, with a peak luminance of 410 lux. The viewing distance was set to 121cm (1.5 times the screen height) for Resolution Group A (UHD) and 241cm (three times the screen height) for Resolution Group B (HD) and C (HD-DO), following the

²We have not compared VVC with other codecs using the DO approach. This is mainly due to the high computation complexity of VVC and the limited computational resources that we have. Preliminary results for Group A and B have already shown the significant improvement of VVC over the other two.

TABLE II: Pre-determined target bit rates for all test sequences in three resolution groups.

Sequence	Target Bitrates (kbps)													
	Resolution Group A (UHD only)				Resolution Group B (HD only)				Resolution Group C (HD-DO)					
	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	R5	
V1: AirAcrobatic	1300	2250	4700	9270	550	920	1850	3400	305	575	940	1770	3350	
V2: CatRobot	3170	5450	8450	14500	1480	2200	3250	5440	910	1500	2200	3250	5500	
V3: Myanmar	11000	21100	33500	46000	3450	5500	8200	10800	2100	3450	5500	8150	10800	
V4: CalmingWater	10100	19250	30000	50000	3100	6400	12000	21000	1140	3050	6550	12200	20500	
V5: ToddlerFountain	13180	27000	38350	69800	6150	13500	21500	34500	2900	5900	13200	20500	34900	
V6: LampLeaves	14550	26460	43900	69800	8100	14200	20500	33000	5030	8100	14000	20500	33500	
V7: DaylightRoad	2650	4450	7050	12170	1220	1800	3000	5300	810	1220	1800	3000	5300	
V8: RedRock	1500	2600	4000	6380	680	1000	1650	2500	460	650	1020	1600	2450	
V9: RollerCoaster	1750	2880	4600	7350	880	1480	2270	3564	550	850	1480	2280	3580	

recommendation in ITU-R BT.500 [27] and ITU-R P.910 [34]. The presentation of video sequences was controlled by a Windows PC running an open source software, BVI-SVQA [35], developed at the University of Bristol for psychophysical experiments.

B. Experimental Procedure

In all three experiments, the Double Stimulus Continuous Quality Scale (DSCQS) [27] methodology was used. In each trial, participants were shown a pair of sequences twice, including original and encoded versions. The presentation order was randomised in each trial and was unknown to each participant. Participants had unlimited time to respond to the following question (presented on the video monitor): “Please rate the quality (0-100) of the first/second video. Excellent–90, Good–70, Fair–50, Poor–30 and Bad–10”. Participants then used a mouse to scroll through the vertical scale and score (0-100) for these two videos. The total duration of each experimental session was approximately 50 (Resolution Group A and B) or 60 (Resolution Group C) minutes, and each was split into two sub-sessions with a 10 minute break in between. Before the formal test, there was a training session consisting of three trials (different from those used in the formal test).

C. Participants and Data Processing

A total of 60 subjects (20 for each test session), with an average age of 27 (age range 20-45), from the University of Bristol were compensated for their participation in the experiments. All of them were tested for normal or corrected-to normal vision. Responses from the subjects were first recorded as quality scores in the range 0-100, as explained earlier. Difference scores were then calculated for each trial and each subject by subtracting the quality score of the distorted sequence from its corresponding reference. Difference Mean Opinion Scores (DMOS) were then obtained for each trial by taking the mean of the difference scores among participants.

V. RESULTS AND DISCUSSION

This section presents the codec comparison results based on objective and subjective quality assessments, alongside

encoder and decoder complexity assessments. For the objective evaluation, two video quality metrics have been employed: the commonly used Peak-Signal-to-Noise-Ratio (PSNR) and Video Multi-method Assessment Fusion (VMAF) [33]. The latter is a machine learning-based video quality metric, which predicts subjective quality by combining multiple quality metrics and video features, including the Detail Loss Metric (DLM) [36], Visual Information Fidelity measure (VIF) [37], and averaged temporal frame difference [33]. The fusion process employs a ν -Support Vector machine (ν -SVM) regressor [38]. VMAF has been evaluated on various video quality databases, and shows improved correlation with subjective scores [33, 39, 40]. In this work VMAF has also been employed to determine optimum resolution for each test rate point and sequence, following the procedure described in Section III-B. The difference between test video codecs in terms of coding efficiency was calculated using the Bjøntegaard Delta (BD) measurements [20] benchmarked against HEVC HM.

The rate-quality curves are plotted for each test sequence in all three resolution groups, where the subjective quality is defined as 100-DMOS for each rate point. A significance test was then conducted using one-way Analysis of Variance (ANOVA) [41, 42] between each paired of codecs on all rate points and sequences.

The subjective data has also been used to evaluate six popular objective video quality metrics, including PSNR, Structural Similarity Index (SSIM) [43], multi-scale SSIM (MS-SSIM) [44], VIF [37], Visual Signal-to-Noise Ratio (VSNR) [45], and VMAF. Following the procedure in [46], their quality indices and the subjective DMOS were fitted based on a weighted least-squares approach using a logistic fitting function for three different resolution groups. The correlation performance of these quality metrics was assessed using four correlation statistics, the Spearman Rank Order Correlation Coefficient (SROCC), the Linear Correlation Coefficient (LCC), the Outlier Ratio (OR) and the Root Mean Squared Error (RMSE). The definitions of these parameters can be found in [46, 47].

Finally, the computational complexity of the three tested encoders was calculated and normalised to HEVC HM for Resolution Group A and B. They were executed on the CPU

nodes of a shared cluster, Blue Crystal Phase3 [48] based at the University of Bristol. Each node has 16×2.6 GHz SandyBridge cores and 64GB RAM.

A. Results based on objective quality assessment

Table III summarises the Bjøntegaard Delta measurements (BD-rate) [20] of AOM AV1 (for three resolution groups) and VVC VTM (for Resolution Group A and B only) compared with HEVC HM, based on both PSNR and VMAF. For the tested codec versions and configurations, it can be observed that AV1 achieves an average bit rate saving of 7.3% against HEVC HM for the UHD test content assessed by PSNR, and this figure reduces (3.8%) at HD resolution. When VMAF is employed for quality assessment, the coding gains of AV1 over HM are slightly higher, averaging 8.6% and 5.0% for UHD and HD respectively. Comparing to AV1, VTM provides significant bit rate savings for both HD and UHD test content, with average BD-rate values between -27% and -30% for PSNR and VMAF. For resolution group C, where VMAF-based DO was applied for HM and AV1, the coding gain achieved by AV1 is 6.3% (over HM) assessed by VMAF, while there is a BD-rate (1.8%) loss when PSNR is employed. In overall conclusion, the performance of AV1 makes a small improvement over HM on the test content, and both AV1 and HM perform (significantly) worse than VTM.

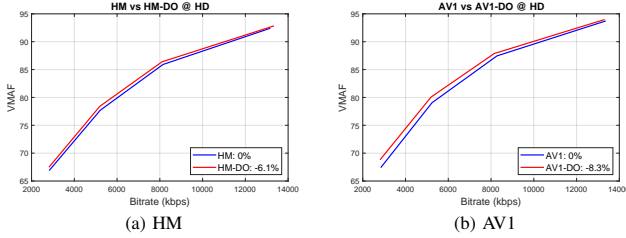


Fig. 3: The average rate-VMAF curves of the nine test sequences for HM and AV1 with and without applying DO.

In order to further compare performance in the context of dynamic optimization (DO), the average rate-VMAF curves of the nine test sequences (HD resolution only) for HM and AV1 with and without DO are shown in Fig. 3. It can be observed that DO has achieved slightly higher overall coding gains for AV1 (BD-rate is -8.2%) on the tested content compared to HM (BD-rate is -6.1%). For both codecs, the savings become lower for higher bitrates (low QP and high quality). It should be noted that the DO approach employed was based on efficient up-sampling using simple spatial filters. More significant improvement has been reported [49, 50] when more advanced up-sampling approaches are applied, such as deep learning based super-resolution.

B. Results based on subjective quality assessment

For each resolution group, we performed outlier rejection on all participant scores. No participants were rejected. Then, we performed one-way ANOVA analysis between pairs of the tested codecs to assess the significance of the differences. Tables IV-VI summarise this comparison.

1) *Resolution Group A (UHD)*: As can be seen in Table IV, the significance tests indicate only two points which exhibit significant difference ($p < .05$) between HM and AV1: Myanmar sequence at R1 and LampLeaves at R2. The reason that significant differences are noticed in the case of the Myanmar sequence may be associated with the observation that, at lower bit rates, the AV1 encoder demonstrates noticeable artifacts on regions of interest (on the heads of walking monks). Performing significance tests between VTM and HM, more cases with significant difference were identified. Particularly, for CatRobot at R1 and R3; for LampLeaves at R1; for DaylightRoad at R1; and for RedRock at R1. Similarly, the significant differences between VTM and AV1 are: for CatRobot at R1; for Myanmar at R1; for DaylightRoad at R2 and R3; and for RedRock at R1.

2) *Resolution Group B (HD)*: Generally, the results for the resolution Group B align with those for Group A. We performed outlier rejection on the participant scores and, again, no one was rejected. We then performed one-way ANOVA analysis between pairs of the tested codecs to assess the significance of the differences. Table V summarises this comparison. From this Table, it can be observed that VTM is significantly better than HM in 14 cases and better than AV1 in 15 cases, and these figures are a factor of three times more than those for the 4K results.

3) *Resolution Group C (HD-DO)*: After performing the significance test using one-way Analysis of Variance (ANOVA) between paired AV1 and HM sequences, only six rate points were indicated as significantly different. This is illustrated by the p -values for each of the 45 encoded pairs of AV1/HEVC sequences. In six cases, HM is significantly better than AV1 ($p < 0.05$): at R4 for AirAcrobatic; at R1-R4 for Myanmar; and at R2 for LampLeaves.

C. Objective Quality Metric Performance Comparison

The correlation performance of six tested objective quality metrics for three resolution groups (in terms of SROCC values) is summarised in Table VII. It can be observed that VMAF outperforms the other five metrics on all three test databases with the highest SROCC and LCC values, and lowest OR and RMSE. PSNR results in much lower performance, especially for the UHD resolution group. It is also noted that, for all test quality metrics, the SROCC values for three resolution groups are all below 0.9, which indicates that further enhancement is still needed to achieve more accurate prediction.

D. Computational complexity analysis

The average complexity figures for encoding UHD and HD content are summarised in Table VIII, where the HM encoder has been used for benchmarking. The average complexity is computed as the average ratio of the execution time of the tested codec for all rate points over the benchmark. As can be seen, for the tested codec versions, AV1 has a higher complexity compared to VTM³. Interestingly these figures are

³It is noted that the complexity for AV1 in more recent versions have been significantly reduced.

TABLE III: Codec comparison results based on PSNR and VMAF quality metrics. Here Bjøntegaard Delta [20] measurements (BD-rate) were employed, and HEVC HM was used as benchmark.

Resolution Group	A (UHD)				B (HD)				C (HD-DO)	
Codec	PSNR		VMAF		PSNR		VMAF		PSNR	VMAF
Sequence\BD-rate	AV1	VTM	AV1	VTM	AV1	VTM	AV1	VTM	AV1	AV1
AirAcrobatic	-12.1%	-25.5%	-12.0%	-28.6%	-2.6%	-21.7%	4.2%	-20.0%	13.3%	-0.1%
CatRobot	-6.2%	-38.0%	-12.8%	-39.6%	-4.0%	-37.7%	-10.4%	-41.2%	-2.1%	-11.3%
Myanmar	4.3%	-17.2%	1.3%	-21.3%	6.5%	-15.5%	3.5%	-18.6%	8.4%	5.1%
CalmingWater	-15.5%	-21.5%	-9.6%	-18.9%	-15.7%	-22.6%	-10.2%	-19.6%	-13.0%	-10.4%
ToddlerFountain	-6.6%	-18.7%	-2.0%	-17.4%	-8.1%	-18.2%	-3.7%	-16.4%	-7.8%	-7.3%
LampLeaves	-6.8%	-26.2%	-6.2%	-26.1%	-2.8%	-23.7%	-0.4%	-24.8%	6.7%	-1.2%
DaylightRoad	-3.8%	-38.0%	-12.4%	-40.3%	-0.9%	-37.6%	-10.4%	-42.4%	0.9%	-9.8%
RedRock	-3.5%	-32.5%	-9.0%	-37.9%	0.8%	-31.5%	-6.4%	-37.7%	16.1%	-8.0%
RollerCoaster	-15.3%	-39.9%	-14.5%	-41.7%	-7.9%	-38.9%	-11.5%	-39.8%	-6.6%	-13.5%
Average	-7.3%	-28.5%	-8.6%	-30.2%	-3.8%	-27.5%	-5.0%	-28.9%	1.8%	-6.3%

TABLE IV: Aggregated significant difference of perceived quality among the tested codecs. The maximum points that can be reached is 36 (aka the number of tested sequences).

Codecs	AV1	HM	VTM
AV1	-	2/36, (1/-1)	5/36, (0/-5)
HM	2/36, (1/-1)	-	5/36, (0/-5)
VTM	5/36, (5/0)	5/36, (5/0)	-

TABLE V: Aggregated significant difference of perceived quality among the tested codecs. The maximum points that can be reached is 36 (aka the number of tested sequences).

Codecs	AV1	HM	VTM
AV1	-	2/36, (0/-2)	15/36, (0/-15)
HM	2/36, (2/0)	-	14/36, (0/-14)
VTM	15/36, (15/0)	14/36, (14/0)	-

higher for the HD than the UHD resolution. The relationship between the relative complexity and encoding performance (in terms of average coding gains for PSNR and VMAF) is also shown in Fig. 4.

VI. CONCLUSIONS

This paper presents performance evaluation results for three major contemporary video codecs, HEVC HM, AV1, and VVC VTM, based on both objective and subjective assessments. Representative test sequences at UHD and HD resolutions were encoded using these codecs to achieve pre-defined target bitrates. The convex hull rate-distortion optimisation has been further employed to compare HEVC HM and AV1 across different resolutions (HD and below) and across a wider bit rate range. The collected subjective data have also been used to evaluate six commonly used quality metrics. Overall, for the tested versions, HM and AV1 are not significantly different in terms of perceived quality at the same bit rates and all resolution-groups. The tested VTM version is however performing significantly better than HM and AV1. All the orig-

TABLE VI: Aggregated significant difference of perceived quality among the tested codecs. The maximum points that can be reached is 45 (aka the number of tested sequences).

Codecs	AV1	HM
AV1	-	6/45, (0/-6)
HM	6/45, (6/0)	-

inal and compressed video sequences and their corresponding subjective scores are now available online⁴ for public testing.

REFERENCES

- [1] CISCO, “CISCO visual networking index: forecast and methodology, 2017–2022,” November 2018.
- [2] ITU-T Rec. H.120, *Codecs for videoconferencing using primary digital group transmission*, ITU-T Std., 1993.
- [3] ITU-T Rec. H.262, *Information technology - Generic coding of moving pictures and associated audio information: Video*, ITU-T Std., 2012.
- [4] ITU-T Rec. H.264, *Advanced Video Coding for Generic Audio-visual Services*, ITU-T Std., 2005.
- [5] B. Bross, J. Chen, S. Liu, and Y.-K. Wang, “Versatile video coding (draft 7),” in the *JVET meeting*, no. JVET-P2001. ITU-T and ISO/IEC, 2019.
- [6] ITU-T Rec. H.265, *High efficiency video coding*, ITU-T Std., 2015.
- [7] AOM. (2019) AOMedia Video 1 (AV1). [Online]. Available: <https://github.com/AOMediaCodec>
- [8] VP9 Video Codec. Google. [Online]. Available: <https://www.webmproject.org/vp9/>
- [9] P. Akyazi and T. Ebrahimi, “Comparison of compression efficiency between HEVC/H. 265, VP9 and AV1 based on subjective quality assessments,” in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.
- [10] D. Grois, T. Nguyen, and D. Marpe, “Coding efficiency comparison of AV1, VP9, H.265/MPEG-HEVC, and H. 264/MPEG-AVC encoders,” in *Picture Coding Symposium (PCS)*. IEEE, 2016, pp. 1–5.

⁴<https://vilab.blogs.bristol.ac.uk/?p=2295>

TABLE VII: The correlation statistics of six popular quality metrics when evaluated on three subject datasets (UHD, HD and HD-DO).

Database Metric	UHD (108)				HD (108)				HD-DO (90)			
	SROCC	LCC	OR	RMSE	SROCC	LCC	OR	RMSE	SROCC	LCC	OR	RMSE
PSNR	0.5517	0.6278	0.3056	8.7540	0.6097	0.6268	0.5556	12.5870	0.7462	0.7439	0.4222	13.3191
SSIM	0.5911	0.5853	0.3148	9.2195	0.7194	0.6757	0.4907	11.5968	0.8026	0.7836	0.3778	12.2184
MSSSIM	0.7426	0.7436	0.2130	7.4102	0.7534	0.7241	0.4537	10.7594	0.8321	0.8228	0.3556	11.1398
VIF	0.7464	0.7749	0.1852	6.9273	0.7459	0.7592	0.3796	10.0815	0.8232	0.8321	0.3778	10.8851
VSNR	0.5961	0.6580	0.2500	8.4062	0.5763	0.6587	0.3889	12.0502	0.6581	0.7039	0.4778	14.0736
VMAF	0.8463	0.8375	0.1574	5.9972	0.8723	0.8476	0.2870	7.9969	0.8783	0.8840	0.2556	9.1395

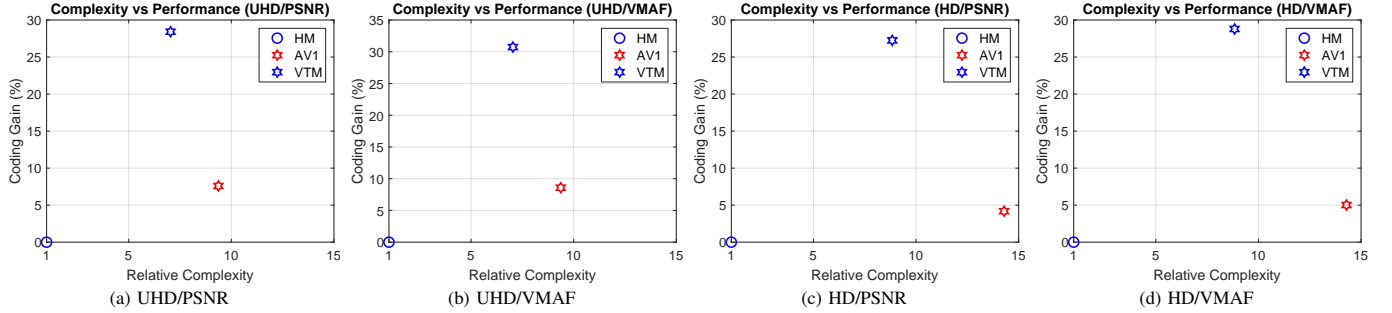


Fig. 4: The relationship between the relative codec complexity (benchmarked on HM) and encoding performance (in terms of average coding gains) for different resolution groups and quality metrics.

TABLE VIII: Computational complexity comparison.

Resolution Group/Codecs	HM	AV1	VTM
Resolution Group A (UHD)	1	9.37×	7.04×
Resolution Group B (HD)	1	14.29×	8.84×

- [11] A. S. Dias, S. Blasi, F. Rivera, E. Izquierdo, and M. Mrak, “An overview of recent video coding developments in MPEG and AOMedia,” in *International Broadcasting Convention (IBC)*, 2018.
- [12] L. Guo, J. De Cock, and A. Aaron, “Compression performance comparison of x264, x265, libvpx and aomenc for on-demand adaptive streaming applications,” in *Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 26–30.
- [13] A. Zabrovskiy, C. Feldmann, and C. Timmerer, “A practical evaluation of video codecs for large-scale HTTP adaptive streaming services,” in *25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 998–1002.
- [14] I. Katsavounidis and L. Guo, “Video codec comparison using the dynamic optimizer framework,” in *Applications of Digital Image Processing XLI*, vol. 10752. International Society for Optics and Photonics, 2018, p. 107520Q.
- [15] I. Katsavounidis, “Dynamic optimizer – a perceptual video encoding optimization framework,” *The Netflix Tech Blog*, 2018.
- [16] A. V. Katsenou, F. Zhang, M. Afonso, and D. R. Bull, “A subjective comparison of AV1 and HEVC for adaptive video streaming,” in *Proc. IEEE Int Conf. on Image Processing*, 2019.
- [17] D. R. Bull, *Communicating pictures: a course in image and Video Coding*. Academic Press, 2014.
- [18] M. Wien, *High efficiency video coding*. Springer, 2015.
- [19] J.-R. Ohm, *Multimedia signal coding and transmission*. Springer, 2015.
- [20] G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” in *13th VCEG Meeting*, no. VCEG-M33. Austin, Texas, USA: ITU-T, April 2001.
- [21] P. Hanhart and T. Ebrahimi, “Calculation of average coding efficiency based on subjective quality scores,” *Journal of Visual Communication and Image Representation*, vol. 25, no. 3, pp. 555 – 564, 2014, qoE in 2D/3D Video Systems. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1047320313002095>
- [22] J. S. Lee, F. De Simone, and T. Ebrahimi, “Subjective quality evaluation via paired comparison: application to scalable video coding,” *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.
- [23] K. Sharman and K. Suehring, “Common test conditions for hm video coding experiments,” in *the JCT-VC meeting*, no. JCTVC-AF1100. ITU-T, ISO/IEC, 2018.
- [24] T. Daede, A. Norkin, and I. Brailovskiy, “Video codec testing and quality measurement,” in *Internet-Draft*. Network Working Group, 2019.
- [25] F. Bossen, J. Boyce, X. Li, V. Seregin, and K. Sühring, “JVET common test conditions and software reference configurations for SDR video,” in *the JVET meeting*, no. JVET-M1001. ITU-T and ISO/IEC, 2019.
- [26] Harmonic, “Harmonic free 4K demo footage,” <https://www.harmonicinc.com/news-insights/blog/4k-in-context/>.
- [27] Recommendation ITU-R BT.500-12, *Methodology for the subjective assessment of the quality of television pictures*, ITU-R Std., 2012.
- [28] F. Mercer Moss, K. Wang, F. Zhang, R. Baddeley, and D. R. Bull, “On the optimal presentation duration for subjective video quality assessment,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 1977–1987, 2016.
- [29] F. Mercer Moss, C.-T. Yeh, F. Zhang, R. Baddeley, and D. R. Bull, “Support for reduced presentation durations in subjective video quality assessment,” *Signal Processing: Image Communication*, vol. 48, pp. 38–49, 2016.
- [30] F. Zhang and D. R. Bull, “A parametric framework for video compression using region-based texture models,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1378–1392, 2011.
- [31] S. Winkler, “Analysis of public image and video database for

- quality assessment,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 1–10, 2012.
- [32] C. E. Duchon, “Lanczos filtering in one and two dimensions,” *Journal of applied meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.
- [33] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward a practical perceptual video quality metric,” *The Netflix Tech Blog*, 2016.
- [34] P.910, *Subjective video quality assessment methods for multimedia applications*, ITU-T Std. Recommendation ITU-T P.910, 1999.
- [35] G. Dimirov, A. V. Katsenou, and D. R. Bull, “SVQA: Subjective Video Quality Assessment software,” <https://github.com/goceee/SVQA>, Aug. 2019.
- [36] S. Li, F. Zhang, L. Ma, and K. H. Ngan, “Image quality assessment by separately evaluating detail losses and additive impairments,” *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.
- [37] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” *IEEE Transactions on Image Processing*, vol. 14, pp. 2117–2128, 2005.
- [38] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [39] F. Zhang, F. Mercer Moss, R. Baddeley, and D. R. Bull, “BVI-HD: A video quality database for HEVC compressed and texture synthesised content,” *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, October 2018.
- [40] J. Li, L. Krasula, Y. Baveye, Z. Li, and P. Le Callet, “Accann: A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience,” *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2589–2602, 2019.
- [41] M. Narwaria, L. Krasula, and P. Le Callet, “Data analysis in multimedia quality assessment: Revisiting the statistical tests,” *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2063–2072, 2018.
- [42] A. V. Katsenou, G. Dimitrov, D. Ma, and D. Bull, “BVI-SynTex: A synthetic video texture dataset for video compression and quality assessment,” *IEEE Transactions on Multimedia*, 2020.
- [43] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, 2004.
- [44] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in *Proc. Asilomar Conference on Signals, Systems and Computers*, vol. 2. IEEE, 2003, p. 1398.
- [45] D. Chandler and S. Hemami, “VSNR: A wavelet-based visual signal-to-noise ratio for natural images,” *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.
- [46] Video Quality Experts Group, “Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment.” 2000. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phase1
- [47] F. Zhang and D. R. Bull, “A perception-based hybrid model for video quality assessment,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2016.
- [48] University of Bristol, “BlueCrystal Phase 3.” [Online]. Available: <https://www.acrc.bris.ac.uk/acrc/phase3.htm>
- [49] M. Afonso, F. Zhang, and D. R. Bull, “Video compression based on spatio-temporal resolution adaptation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 275–280, January 2019.
- [50] F. Zhang, M. Afonso, and D. R. Bull, “Vistra2: Video coding using spatial resolution and effective bit depth adaptation,” *arXiv preprint arXiv:1911.02833*, 2019.



Fan Zhang (M’12) received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University (2005 and 2008 respectively), and his Ph.D from the University of Bristol (2012). He is currently working as a Research Assistant in the Visual Information Laboratory, Department of Electrical and Electronic Engineering, University of Bristol, on projects related to perceptual video compression. His research interests include perceptual video compression, video quality assessment and immersive video formats including HDR and HFR.



Angeliki V. Katsenou is a Leverhulme Early Career Fellow and is with the Visual Information Lab at the University of Bristol, U.K., since 2015. She obtained her Ph.D. degree from the Department of Computer Science and Engineering, University of Ioannina, Greece (2014). She received her M.Eng. in Electrical and Computer Engineering and the M.Sc. degree in Signal and Image Processing from the University of Patras, Greece. She has experience in several EC-funded and EPSRC projects, such as MSCA-ITN PROVISION and EPSRC Platform Grant EP/M000885/1. Her research interests include perceptual video analysis, video compression, and quality.



Mariana Afonso received the B.S./M.S. degree in Electrical and Computers Engineering from the University of Porto, Portugal, in 2015 and a PhD in Electronic and Electrical Engineering from the University of Bristol, U.K., in 2019. During her PhD, she was a part of MSCA-ITN PROVISION, a network of leading academic and industrial organizations in Europe, working on perceptual video coding. She also completed a secondment at Netflix, USA, in 2017. She is currently a Research Scientist in the Video Algorithms team at Netflix. Her research interests include video compression, video quality assessment, and machine learning.



Goce Dimitrov received his B.Sc./M.Eng. in Computer Science and Electronics at the University of Bristol, U.K., in 2019. During his final year project he developed a software for subjective video quality assessment and he did an internship on video codec comparison with the Visual Information Lab, University of Bristol, under the supervision of Prof. D. Bull and Dr. A. Katsenou. His research interests include software engineering, video quality assessment, cloud computing and artificial intelligence.



David R. Bull (M'94-SM'07-F'12) received the B.Sc. degree from the University of Exeter, Exeter, U.K., in 1980; the M.Sc. degree from University of Manchester, Manchester, U.K., in 1983; and the Ph.D. degree from the University of Cardiff, Cardiff, U.K., in 1988.

Dr Bull has previously been a Systems Engineer with Rolls Royce, Bristol, U.K. and a Lecturer at the University of Wales, Cardiff, U.K. He joined the University of Bristol in 1993 and is currently its Chair of Signal Processing and Director of Bristol Vision Institute. In 2001, he co-founded a university spin-off company, ProVision Communication Technologies Ltd., specializing in wireless video technology. He has authored over 450 papers on the topics of image and video communications and analysis for wireless, Internet and broadcast applications, together with numerous patents, several of which have been exploited commercially. He has received two IET Premium awards for his work. He is the author of three books, and has delivered numerous invited/keynote lectures and tutorials. Dr. Bull is a fellow of the Institution of Engineering and Technology.